

Toward computerized efficient estimation in infinite-dimensional models

Marco Carone¹, Alexander R. Luedtke² and Mark J. van der Laan³

¹Department of Biostatistics, University of Washington

²Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center

³Division of Biostatistics, University of California, Berkeley

September 1, 2016

Abstract

Despite the risk of misspecification they are tied to, parametric models continue to be used in statistical practice because they are accessible to all. In particular, efficient estimation procedures in parametric models are simple to describe and implement. Unfortunately, the same cannot be said of semiparametric and nonparametric models. While the latter often reflect the level of available scientific knowledge more appropriately, performing efficient inference in these models is generally challenging. The efficient influence function is a key analytic object from which the construction of asymptotically efficient estimators can potentially be streamlined. However, the theoretical derivation of the efficient influence function requires specialized knowledge and is often a difficult task, even for experts. In this paper, we propose and discuss a numerical procedure for approximating the efficient influence function. The approach generalizes the simple nonparametric procedures described recently by Frangakis et al. (2015) and Luedtke et al. (2015) to arbitrary models. We present theoretical results to support our proposal, and also illustrate the method in the context of two examples. The proposed approach is an important step toward automating efficient estimation in general statistical models, thereby rendering the use of realistic models in statistical analyses much more accessible.

Keywords: asymptotic efficiency, canonical gradient, efficient influence function, infinite-dimensional models, pathwise differentiability.

Corresponding author: Marco Carone, Department of Biostatistics, University of Washington, F-644, Health Sciences Building, Box 357232, Seattle, WA 98195-7232. Email: mcarone@uw.edu.

1 Introduction

Efficient estimation techniques are often preferred because they maximally exploit available information and minimize the uncertainty of the resulting scientific findings. Efficiency is most broadly defined in an asymptotic sense. As such, characterizing asymptotic efficiency and constructing asymptotically efficient estimators has been an important focus of methodological and theoretical research in statistics. For convenience, throughout this paper, we ascribe an asymptotic sense to the terms *efficient* and *efficiency*.

In the context of parametric models, a simple efficiency theory has been available for nearly a century, largely established in Fisher’s work on maximum likelihood estimation. In such models, efficiency is characterized by the Cramer-Rao bounds and efficient estimators can generally be obtained via maximum likelihood (see, e.g., Hájek, 1970, 1972; Le Cam, 1972). When parametric models are adopted in practice, it is often because they are simple and convenient to use. However, the use of such models carries the potential for model misspecification, which may have potentially serious adverse effects on the scientific process. In many scientific problems, the available background knowledge simply does not justify the use of such restrictive statistical models.

Infinite-dimensional models – either nonparametric or semiparametric – offer a more flexible alternative. These richer models mitigate the risk of model misspecification and more accurately reflect the level of available prior knowledge. Unfortunately, establishing efficiency bounds for target parameters in infinite-dimensional models can be a very complex task. The development of a general efficiency theory, valid for arbitrary statistical models, is a more recent accomplishment: except for the early seminal contribution of Stein (1956), developments in this area began in the late 1970s and early 1980s with the works of Koshevnik and Levit (1977), Pfanzagl (1982) and Begun et al. (1983), among others, and continued throughout the 1990s (see, e.g., van der Vaart, 1991; Newey, 1994). Notably, it builds upon notions of differential geometry and functional analysis. In certain cases, a generalized notion of maximum likelihood, as described, for example, by Kiefer and Wolfowitz (1956), can still be used to produce efficient estimators. In other cases though, the statistical model is too complex for a maximum likelihood estimator to exist, let alone be well-behaved. This renders the pursuit of efficient estimators a substantially more difficult task in infinite-dimensional models.

A key object in this general efficiency theory is the efficient influence function, hereafter referred to as EIF. It bears this name because it is the influence function of any efficient estimator of the parameter of interest given a particular statistical model. If the EIF is known, efficiency bounds can easily

be estimated, at least theoretically, and the performance of candidate estimators can be examined against an objective benchmark. Valid confidence intervals based on a given efficient estimator can also be constructed using the EIF. This is particularly useful in settings where the bootstrap is known to fail. More importantly, if the analytic form of the EIF is available, efficient estimators can be constructed rather easily. To do so, several approaches may be used, including, for example, gradient-based estimating equations (e.g., van der Laan and Robins, 2003), Newton-Raphson one-step corrections (e.g., Pfanzagl, 1982) and targeted minimum loss-based estimation (e.g., van der Laan and Rose, 2011). This provides a strong motivation for deriving the EIF in a given statistical problem. Unfortunately, the analytic computation of the EIF is seldom straightforward. It generally involves finding an influence function, characterizing the tangent space of the statistical model and projecting onto it – the effort can be mathematically intricate. Over the years, many techniques have been developed to facilitate this task in certain classes of problems – the discretization technique of Chamberlain (1987) is one such example. Despite this, this calculation remains a rather specialized skill, mastered mostly by a small collection of theoretically-inclined researchers. The theoretical derivation of EIFs is generally not in the skill set of practicing statisticians. Yet, in many problems, it is a necessary skill to master in order to make optimal inference in more realistic statistical models. The paucity of this skill has likely constituted an impediment to a broader appreciation and adoption of semiparametric and nonparametric techniques in applications.

In view of this barrier, one naturally wonders whether a suitable numerical approximation could serve as substitute for the analytic form of the EIF, and further whether its calculation could be computerized. An affirmative answer to this question would render the implementation of efficient inferential techniques in semiparametric and nonparametric models much more accessible to practitioners, and the impact on current statistical practice could be profound. Recently, a very important first step toward this goal was made by Frangakis et al. (2015): these authors proposed a simple numerical routine for calculating the EIF in the context of nonparametric models when the data are discrete-valued or when the parameter is a smooth functional of the distribution function. In our discussion of their article (see Luedtke et al., 2015), we suggested a regularization of their technique that is valid more broadly within the context of nonparametric models. Nevertheless, neither of these methods formally address the more difficult problem of computerizing the calculation of the EIF in semiparametric models. As opposed to nonparametric models, for which the tangent space is trivially described, semiparametric models generally have much more complex tangent spaces, projecting onto

which may often require great skill. Identifying a numerical approach for computing the EIF in semi-parametric models is therefore a more difficult but also more needed innovation. In this article, we establish and study novel representations of the EIF that naturally allows a numerical computation of the EIF of a given parameter in a given statistical model. Importantly, we do not impose constraints on the type of model that may be considered. These representations hold great promise in allowing true computerization, as we discuss below.

This paper is organized as follows. In Section 2, we present novel representations of the EIF for use in arbitrary statistical models and show how they may be used to calculate the EIF numerically. In Section 3, we establish sufficient technical conditions that guarantee the validity of these representations. We discuss various practical issues regarding the implementation of our proposal in Section 4. In Section 5, we illustrate the validity and feasibility of the approach in the context of two examples. Finally, we provide concluding remarks in Section 6. While Theorem 1 is proved in the body of the paper, the proof of Theorems 2, 3 and 4 are provided in an Appendix.

2 Numerical calculation of the efficient influence function

2.1 Preliminaries

Suppose that we observe independent d -dimensional variates X_1, X_2, \dots, X_n following a distribution P_0 known only to belong to the statistical model \mathcal{M} . We denote by $\mathcal{X}(P) \subseteq \mathbb{R}^d$ the sample space associated to $P \in \mathcal{M}$. We are interested in efficiently inferring about $\psi_0 := \Psi(P_0)$ using the available data, where $\Psi : \mathcal{M} \rightarrow \mathbb{R}^q$ represents a pathwise differentiable parameter mapping of interest. Pathwise differentiability ensures the parameter is a sufficiently smooth mapping so as to admit an efficiency theory (see, e.g., Pfanzagl, 1982; Bickel et al., 1997). We denote by $L_2^0(P)$ the Hilbert space of P -integrable functions from $\mathcal{X}(P)$ to \mathbb{R}^q with mean zero and finite variance under P . The parameter Ψ is said to be pathwise differentiable if there exists some $\chi_P \in L_2^0(P)$ such that, for each regular one-dimensional parametric submodel $\mathcal{M}_0 := \{P_\epsilon : \epsilon \in \mathcal{E}\} \subseteq \mathcal{M}$ with $\mathcal{E} \subset \mathbb{R}$ an interval containing zero and $P_{\epsilon=0} = P$, the pathwise derivative $\left. \frac{d}{d\epsilon} \Psi(P_\epsilon) \right|_{\epsilon=0}$ can be represented as the inner product $\int \chi_P(u) s(u) dP(u)$, where s is the score for ϵ at $\epsilon = 0$ in \mathcal{M}_0 (Pfanzagl, 1982). Any such element χ_P is said to be a gradient of Ψ at P relative to \mathcal{M} . The tangent space $T_{\mathcal{M}}(P)$ of \mathcal{M} at P is defined as the closure of the linear span of scores at P arising from regular one-dimensional parametric submodels of \mathcal{M} through P . The canonical gradient is the unique gradient contained in $T_{\mathcal{M}}(P)$ and corresponds

to the EIF under sampling from P . Throughout, we will refer to the EIF at P as ϕ_P and write $\phi_P(x)$ for the evaluation of ϕ_P at the observation value x . The asymptotic variance of an efficient estimator of ψ_0 relative to model \mathcal{M} is given by $\int \phi_{P_0}(u)\phi_{P_0}(u)^\top dP_0(u)$. Without loss of generality, we will assume $q = 1$ since the general case can be trivially dealt with using the developments herein applied to each component.

If pathwise differentiability holds uniformly over paths in a neighborhood around P , for any $P_1 \in \mathcal{M}$ close enough to P , the parameter admits the linearization

$$\begin{aligned}\Psi(P_1) - \Psi(P) &= \int \phi_{P_1}(u)d(P_1 - P)(u) + R(P_1, P) \\ &= - \int \phi_{P_1}(u)dP(u) + R(P_1, P)\end{aligned}\tag{1}$$

where $R(P_1, P)$ is a second-order remainder term, and the second line follows from the first since $\int \phi_{P_1}(u)dP_1(u) = 0$ in view of the fact that the EIF is centered. This representation, which is no more than a first-order Taylor approximation over the model space, holds for most smooth parameters arising in practice. The precise form of R is generally established by hand on a case-by-case basis. This linearization is critical for motivating and studying the use of both Newton-Raphson one-step correction and targeted minimum loss-based estimation to construct efficient estimators. It is also at the heart of our current proposal for obtaining a numerical approximation to the EIF value $\phi_P(x)$ at a given distribution $P \in \mathcal{M}$ and observation value $x \in \mathcal{X}(P)$.

2.2 Nonparametric models

Recently, Frangakis et al. (2015) presented one such proposal based on the representation of $\phi_P(x)$ as the Gâteaux derivative of Ψ at P in the direction of $\delta_x - P$, where δ_x represents the degenerate distribution at x . Of course, this can also be seen as the pathwise derivative $\left.\frac{d}{d\epsilon}\Psi(P_\epsilon)\right|_{\epsilon=0}$ of Ψ at P along the linear perturbation path $\{P_\epsilon := (1 - \epsilon)P + \epsilon\delta_x : 0 \leq \epsilon \leq 1\}$ between P and δ_x – this simple observation will be helpful when dealing with arbitrary models. Here and throughout, any such derivative is of course interpreted as a right derivative. To computerize the process of calculating $\phi_P(x)$, these authors suggested approximating this derivative by the slope of the secant line connecting $(0, \Psi(P))$ and $(\epsilon, \Psi(P_\epsilon))$ for a very small $\epsilon > 0$. In our discussion of Frangakis et al. (2015) (see Luedtke et al., 2015), we pointed out sufficient conditions that guarantee that this indeed approximates $\phi_P(x)$. For example, this approach is valid whenever the model \mathcal{M} is nonparametric and the sample space

$\mathcal{X}(P)$ is finite. However, if the parameter Ψ depends on local features of the distribution, this method may fail when $\mathcal{X}(P)$ is infinite, such as when any component of X is continuous under P . We proposed a slight modification of the procedure of Frangakis et al. (2015) to remedy this limitation. Specifically, we proposed replacing the degenerate distribution δ_x at x by a distribution $H_{x,\lambda}$ symmetric about x , dominated by P and such that $\int g(u)dH_{x,\lambda}(u) \rightarrow g(x)$ as $\lambda \rightarrow 0$ for all g in a sufficiently large class of functions. This amounts to replacing the degenerate distribution by a nearly degenerate distribution with smoothing parameter $\lambda > 0$. In a technical report published contemporaneously, Ichimura and Newey (2015) also suggested this approach. As stated in Luedtke et al. (2015), under certain regularity conditions and provided \mathcal{M} is nonparametric, it is generally the case that

$$\phi_P(x) = \lim_{\lambda \rightarrow 0} \frac{d}{d\epsilon} \Psi(P_{\epsilon,\lambda}) \quad (2)$$

$$= \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \frac{\Psi(P_{\epsilon,\lambda}) - \Psi(P)}{\epsilon}, \quad (3)$$

where we have defined the linear perturbation path $P_{\epsilon,\lambda} := (1 - \epsilon)P + \epsilon H_{x,\lambda}$. Representation (2) is useful when the parameter is simple enough so that calculating the derivative of $\epsilon \mapsto \Psi(P_{\epsilon,\lambda})$ is analytically convenient. Otherwise, representation (3) can be used to circumvent this analytic step by approximating this derivative by the slope of a secant line, as in Frangakis et al. (2015). Because these representations constitute a special case of the general result described in the next subsection, we defer a statement of regularity conditions and a formal proof until then.

In practice, to approximate $\phi_P(x)$ numerically, the secant line slope exhibited in (3) is evaluated for small ϵ and λ . This operation only requires the ability to evaluate Ψ on a given distribution. Generally, as we highlighted in Luedtke et al. (2015), ϵ must be chosen much smaller than λ to obtain an accurate approximation – this emphasizes that the order of the limits in (3) plays an important role in the implementation of this procedure. We discuss this point in greater detail later.

2.3 Arbitrary models

When the model is not nonparametric, the representations provided in (2) and (3) generally do not hold. Except for when $\epsilon = 0$, the linear perturbation path described by $P_{\epsilon,\lambda}$ is usually not contained in the model. Therefore, the parameter may not even be defined on this path. Even if it is, in general, the approximation suggested by these representations will at best yield the EIF of Ψ relative to a nonparametric model rather than the actual model. While the EIF in a nonparametric model is still

an influence function in \mathcal{M} , it is not typically efficient. This is also clear from a practical perspective: since the expressions in (2) and (3) do not acknowledge constraints implied by \mathcal{M} , they could not possibly yield the actual EIF.

Since the path $\{P_{\epsilon,\lambda} : 0 \leq \epsilon \leq 1\}$ is generally not in \mathcal{M} for $\epsilon \neq 0$, it appears natural to consider the behavior of Ψ along the analogue of the linear perturbation path in \mathcal{M} . To formalize this idea, we may consider the path $P_{\epsilon,\lambda}^*$ obtained by projecting $P_{\epsilon,\lambda}$ according to the Kullback-Leibler divergence into \mathcal{M} or a suitably regularized version thereof. We formally define

$$P_{\epsilon,\lambda}^* := \operatorname{argmax}_{P_1 \in \mathcal{M}(P)} \int \log \left\{ \frac{dP_1}{d\nu}(x) \right\} dP_{\epsilon,\lambda}(x) , \quad (4)$$

where $\mathcal{M}(P) := \{P_1 \in \mathcal{M} : P_1 \ll P\} \subseteq \mathcal{M}$ is the subset of all probability measures in \mathcal{M} that are absolutely continuous with respect to P , ν is a measure dominating P , and for any $P_1 \in \mathcal{M}(P)$, $dP_1/d\nu$ is the density of P_1 relative to ν . This projection defines a novel path in the model space. Under regularity conditions, we can then establish that (2) and (3) hold more broadly when the linear perturbation path is replaced by the model-specific path defined by this projection, as is formalized below. Here and throughout, we use the shorthand notation $\phi_{\epsilon,\lambda}^*$ to denote $\phi_{P_{\epsilon,\lambda}^*}$.

Theorem 1. *Suppose that $P_{\epsilon,\lambda}^*$ exists and is in \mathcal{M} for all sufficiently small ϵ and λ . Then, provided*

$$(A1) \text{ (solution of EIF estimating equation) } \int \phi_{\epsilon,\lambda}^*(u) dP_{\epsilon,\lambda}(u) = 0;$$

$$(A2) \text{ (continuity of EIF) } \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \int \phi_{\epsilon,\lambda}^*(u) d(H_{x,\lambda} - P)(u) = \phi_P(x);$$

$$(A3) \text{ (preservation of rate of convergence) } \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} R(P_{\epsilon,\lambda}^*, P)/\epsilon = 0;$$

the EIF of Ψ relative to \mathcal{M} at $P \in \mathcal{M}$ evaluated at observation value x is given by

$$\phi_P(x) = \lim_{\lambda \rightarrow 0} \frac{d}{d\epsilon} \Psi(P_{\epsilon,\lambda}^*) \quad (5)$$

$$= \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \frac{\Psi(P_{\epsilon,\lambda}^*) - \Psi(P)}{\epsilon} . \quad (6)$$

Proof. Setting $P_1 = P_{\epsilon,\lambda}^*$ in (1), we note that

$$\begin{aligned} \Psi(P_{\epsilon,\lambda}^*) - \Psi(P) &= - \int \phi_{\epsilon,\lambda}^*(u) dP(u) + R(P_{\epsilon,\lambda}^*, P) \\ &= \int \phi_{\epsilon,\lambda}^*(u) d(P_{\epsilon,\lambda} - P)(u) - \int \phi_{\epsilon,\lambda}^*(u) dP_{\epsilon,\lambda}(u) + R(P_{\epsilon,\lambda}^*, P) . \end{aligned}$$

In view of (A1), we have that $\Psi(P_{\epsilon,\lambda}^*) - \Psi(P) = \int \phi_{\epsilon,\lambda}^*(u) d(P_{\epsilon,\lambda} - P)(u) + R(P_{\epsilon,\lambda}^*, P)$ and since

$P_{\epsilon,\lambda} - P = \epsilon(H_{x,\lambda} - P)$, we find that

$$\frac{\Psi(P_{\epsilon,\lambda}^*) - \Psi(P)}{\epsilon} = \int \phi_{\epsilon,\lambda}^*(u) d(H_{x,\lambda} - P)(u) + \frac{R(P_{\epsilon,\lambda}^*, P)}{\epsilon}.$$

The result follows directly from (A2) and (A3). □

Condition (A1) drives in large part our generalization of the procedures of Frangakis et al. (2015) and Luedtke et al. (2015) to arbitrary models. Projecting the path $\{P_{\epsilon,\lambda} : 0 \leq \epsilon \leq 1\}$ into \mathcal{M} to obtain $\{P_{\epsilon,\lambda}^* : 0 \leq \epsilon \leq 1\}$ is expected to ensure that the score-like equation described in (A1) is solved. In fact, as we will see in the next section, mild regularity conditions ensure that (A1) is satisfied. Condition (A2) imposes relatively weak continuity requirements on the EIF. Since R is a second-order term, $R(P_{\epsilon,\lambda}, P)$ is generally of order $O(\epsilon^2)$ for each fixed $\lambda > 0$. Condition (A3) requires that $R(P_{\epsilon,\lambda}^*, P)$ be of order $o(\epsilon)$ for λ small and ϵ sufficiently smaller. Determining how the projection step and the smoothing parameter $\lambda > 0$ affects the rate of this second-order remainder term is critical to establishing whether (A3) holds. This is studied in detail in the next section.

If the projection $P_{\epsilon,\lambda}^*$ is available in closed form, (5) suggests that we can calculate $\phi_P(x)$ by analytically computing the pathwise derivative of $\epsilon \mapsto \Psi(P_{\epsilon,\lambda}^*)$ at $\epsilon = 0$ and evaluating it at some small value of $\lambda > 0$. If $P_{\epsilon,\lambda}^*$ is not available in closed form or the mapping $\epsilon \mapsto \Psi(P_{\epsilon,\lambda}^*)$ is difficult to differentiate analytically, (6) suggests using the secant line slope

$$\frac{\Psi(P_{\epsilon,\lambda}^*) - \Psi(P)}{\epsilon}$$

for small λ and even smaller ϵ as an approximation to $\phi_P(x)$. Strategies for appropriately selecting values of ϵ and λ are discussed in Section 4, whereas the sensitivity of the approximation to these choices will be studied in the context of two examples in Section 5.

Much of the effort required to use representations (5) and (6) goes into identifying the projection $P_{\epsilon,\lambda}^*$ of $P_{\epsilon,\lambda}$ onto the model space. For this task, the equivalence between minimization of the Kullback-Leibler divergence and maximization of the likelihood is often useful and can be leveraged. In many cases, this projection can be identified analytically. In many others, a numerical approach must be taken. Regardless, the definition of $P_{\epsilon,\lambda}^*$ does not involve the parameter of interest. Hence, the more challenging portion of the approach is exclusively model-specific, and once it has been successfully

tackled, the resulting projection can be used for any parameter a practitioner may wish to study. This contrasts sharply with the conventional approach to deriving the EIF, wherein the statistician must first derive an influence function, characterize the tangent space of the model, and finally project the influence function onto this tangent space. In this conventional approach, both the parameter-specific task – finding an influence function – and the model-specific task – studying the tangent space and how to project onto it – require specialized knowledge. Performing these tasks for a given parameter and model combination does not automatically provide an easy way of tackling any other parameter, in contrast to the approach that we propose.

3 Verification of technical conditions

The validity of representations (5) and (6) is guaranteed to hold under the high-level technical conditions (A1), (A2) and (A3). We now identify lower-level sufficient conditions under which (A1), (A2) and (A3), and thus also Theorem 1, hold.

In the developments below, we let

$$r(\lambda) := \left\| \frac{dH_{x,\lambda}}{dP} \right\|_{2,P} = \sqrt{\int \left\{ \frac{dH_{x,\lambda}}{dP}(u) \right\}^2 dP(u)}$$

denote the $L_2(P)$ -norm of the Radom-Nykodim derivative of $H_{x,\lambda}$ relative to P . This derivative is defined for each $\lambda > 0$ since $H_{x,\lambda}$ is dominated by P by construction. Whenever P does not assign positive mass to the set $\{x\}$, the value of $r(\lambda)$ will usually tend to infinity as λ tends to zero. The rate at which this occurs will be critical in our study of the technical conditions listed in Theorem 1. Here and throughout, given a function h , we define $\|h\|_{2,P} := \sqrt{\int h(u)^2 dP(u)}$ and $\|h\|_{\infty,A} := \sup_{u \in A} |h(u)|$ for any set A . We also denote by $\mathcal{S}_{x,\lambda}$ the support of $H_{x,\lambda}$.

3.1 Solution of the EIF estimating equation

By virtue of being a projection, $P_{\epsilon,\lambda}^*$ is expected to solve a collection of score-like equations, including that exhibited in condition (A1). The following theorem establishes formal regularity conditions validating this heuristic argument.

Theorem 2. *Condition (A1) holds provided either of the following conditions is true:*

- (a) *for some parametric submodel $\mathcal{M}_0 := \{P_\gamma : \gamma \in \Gamma\} \subseteq \mathcal{M}$ where $\Gamma \subseteq \mathbb{R}$ is an interval containing*

zero and $P_{\gamma=0} = P_{\epsilon,\lambda}^*$, the function $u \mapsto \phi_{\epsilon,\lambda}^*(u)$ is the score for γ at $\gamma = 0$ in \mathcal{M}_0 ;

(b) the Radon-Nikodym derivative of $P_{\epsilon,\lambda}$ relative to $P_{\epsilon,\lambda}^*$ is uniformly bounded in $L_2(P_{\epsilon,\lambda}^*)$ -norm.

The tangent space of \mathcal{M} at $P_{\epsilon,\lambda}^*$ is the collection of scores and elements that can be approximated arbitrarily well by a linear combination of scores. Under condition (a) in the above theorem, the result is established automatically since then $\phi_{\epsilon,\lambda}^*$ is itself a score. Condition (b) is a relatively milder condition. It is expected to hold in some generality since $P_{\epsilon,\lambda}^* = P_{\epsilon,\lambda}$ for $\epsilon = 0$ and any $\lambda > 0$, and as such, the Random-Nykodim derivative of $P_{\epsilon,\lambda}$ relative to $P_{\epsilon,\lambda}^*$ equals one at $\epsilon = 0$. Under reasonable continuity, in any small neighborhood of ϵ values near zero, this derivative is expected to be bounded in $L_2(P_{\epsilon,\lambda}^*)$ -norm. Additionally, any region supported by $P_{\epsilon,\lambda}$ and in which $P_{\epsilon,\lambda}^*$ assigns negligible probability mass makes a large negative contribution to the log-likelihood criterion in (4), thereby thwarting the objective of maximizing the likelihood. This observation further supports the plausibility of condition (b), and in fact guarantees it in the context of any finitely-supported $P_{\epsilon,\lambda}$.

3.2 Continuity of the EIF

We relied on certain notions of continuity to establish the validity of representations (5) and (6). The theorem below highlights how the continuity requirement stated in condition (A2) can be more concretely verified.

Theorem 3. *Suppose that $\lim_{\lambda \rightarrow 0} \int \phi_P(u) dH_{x,\lambda}(u) = \phi_P(x)$ and $\lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \int \phi_{\epsilon,\lambda}^*(u) dP(u) = 0$. Condition (A2) holds provided either*

$$(a) \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \|\phi_{\epsilon,\lambda}^* - \phi_P\|_{\infty, \mathcal{S}_{x,\lambda}} = 0 \quad \text{or} \quad (b) \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} r(\lambda) \|\phi_{\epsilon,\lambda}^* - \phi_P\|_{2,P} = 0.$$

The requirement that $\int \phi_P(u) dH_{x,\lambda}(u)$ approximates $\phi_P(x)$ as λ tends to zero simply stipulates that averaging ϕ_P with respect to a distribution eventually concentrating all its probability mass on $\{x\}$ should approximately yield $\phi_P(x)$. Furthermore, this theorem requires that $\int \phi_{\epsilon,\lambda}^*(u) dP(u)$ tends to zero, which is reasonable under some continuity since $\phi_{\epsilon,\lambda}^*$ tends to ϕ_P and $\int \phi_P(u) dP(u) = 0$. Beyond this, in order for condition (A2) to hold, it suffices either for $\phi_{\epsilon,\lambda}^*$ to approximate ϕ_P in supremum norm over the support of $H_{x,\lambda}$ or in $L_2(P)$ -norm at a rate faster than $r(\lambda)^{-1}$. These statements each hinge on a certain notion of continuity that appears needed whenever P is not finitely-supported and nearly degenerate distributions must be used in defining the linear perturbation paths.

3.3 Preservation of the rate of convergence

The proof of representations (5) and (6) hinges upon a linearization of the difference between $\Psi(P_{\epsilon,\lambda}^*)$ and $\Psi(P)$. To ignore the remainder term from this linearization, we require that $R(P_{\epsilon,\lambda}^*, P)/\epsilon$ be arbitrarily small for small enough λ and sufficiently smaller ϵ . The following theorem establishes a bound on $R(P_{\epsilon,\lambda}^*, P)$ in terms of ϵ and λ under mild conditions. It also clarifies how ϵ and λ must be chosen to guarantee condition (A3).

Theorem 4. *Suppose that there exists an interval $I_0 = [m_0, m_1] \subset (0, +\infty)$ such that for each small λ the Radon-Nikodym derivative of $P_{\epsilon,\lambda}^*$ relative to P is uniformly contained in I_0 over the support of P for sufficiently small ϵ . Suppose also that there exist some $0 < C < +\infty$ such that for any $P_1 \in \mathcal{M}$ with Radon-Nikodym relative to P bounded above by m_1 over the support of P we have that*

$$|R(P_1, P)| \leq C \left\| \frac{dP_1}{dP} - 1 \right\|_{2,P}^2.$$

Then, it is true that $R(P_{\epsilon,\lambda}^, P)/[\epsilon\{1+r(\lambda)\}]^2$ is bounded for small λ and sufficiently smaller ϵ . Thus, condition (A3) holds if $\epsilon = \epsilon(\lambda)$ is selected such that $\epsilon(\lambda)\{1+r(\lambda)\}^2 \rightarrow 0$ as λ tends to zero.*

As discussed in the previous subsection, since $P_{\epsilon,\lambda}^* = P$ for any value $\lambda > 0$ whenever $\epsilon = 0$, the derivative of $P_{\epsilon,\lambda}^*$ relative to P is indeed expected to be uniformly bounded above and away from zero for small enough λ and sufficiently smaller ϵ . Furthermore, it is often the case that the remainder term, as being a second-order term arising from a linearization, can be bounded by the squared norm of the difference between the derivative of $P_{\epsilon,\lambda}^*$ relative to P and its value at $\epsilon = 0$. This inequality often follows quite easily from an application of the Cauchy-Schwartz inequality on the remainder term. It is easy to verify in common examples and generally holds under rather mild conditions.

4 Practical considerations

The representations presented in Theorem 1 provide the theoretical foundations for numerically approximating the EIF and thus for numerically constructing efficient estimators. The implementation of the approach suggested by these representations nevertheless presents specific challenges. Practical guidelines, as provided below, may facilitate the successful implementation of our proposal by practitioners.

4.1 Construction of the linear perturbation path

In constructing the linear perturbation path that defines $P_{\epsilon,\lambda}$ and thus $P_{\epsilon,\lambda}^*$, the nearly degenerate distribution at $\{x\}$ is used instead of its purely degenerate counterpart because it ensures that all distributions along the perturbation path are dominated by P . This is required to ensure the validity of the representations we have proposed. Clearly, there is no need for smoothing in the components of the data unit for which the corresponding marginal distribution implied by P is dominated by a counting measure. In fact, as we stress below, unnecessary smoothing will needlessly increase the computational burden of the approximation procedure. For components for which the corresponding marginal distribution is dominated by the Lebesgue measure, smoothing is generally needed. In practice, we suggest the use of product kernels for those components. Specifically, suppose that the data unit X is d -dimensional and can be partitioned into $X = (X_L, X_C)$, where $X_L := (X_{L1}, X_{L2}, \dots, X_{Ld_1})$ and $X_C := (X_{C1}, X_{C2}, \dots, X_{Cd_2})$ with $d_1 + d_2 = d$, and that the marginal distributions of X_L and X_C under P are respectively dominated by the Lebesgue measure and a discrete counting measure. In this case, we can typically use the product kernel

$$u \mapsto H_{x,\lambda}(u) := \left[\prod_{j=1}^{d_1} K_\lambda(u_{Lj} - x_{Lj}) \right] \times \left[\prod_{j=1}^{d_2} I(u_{Cj} = x_{Cj}) \right],$$

where $u := (u_L, u_C)$ with u_L and u_C possible realizations of X_L and X_C , respectively, and $K_\lambda(w) := \lambda^{-1}K(\lambda^{-1}w)$ with K some symmetric, absolutely continuous density function. The uniform kernel $K(w) := I(-1 < 2w < +1)$ is particularly appealing due to its simplicity, which translates to greater practical feasibility of our numerical approximation procedure. If the uniform kernel is used, it is easy to verify that $r(\lambda) = \lambda^{-d_1}$ provided, for example, $u_L \mapsto p(u_L, x_C)$ is continuous and bounded away from zero in a neighborhood of x_L . Thus, to ensure that condition (A3) is satisfied, Theorem 4 suggests choosing ϵ such that $\epsilon \ll \lambda^{2d_1}$. If d_1 is large, this requirement may be prohibitive, possibly even to the point of requiring a value of ϵ beyond the computer's default level of precision and thus requiring special computational techniques. Of course, while this guideline is sufficient, it may be overly conservative in some applications. In the next subsection, we provide a practical means of selecting the value of ϵ and λ .

As alluded to above, if we include smoothing over X_C as well in our choice of $H_{x,\lambda}$, we need $\epsilon \ll \lambda^{2d}$. This can be much more prohibitive computationally than requiring that $\epsilon \ll \lambda^{2d_1}$, particularly if d_2 is large. For this reason, smoothing in the construction of the linear perturbation path should be

avoided for all components except those for which the corresponding marginal distribution under P is absolutely continuous. Additionally, for some parameters, smoothing can be avoided altogether for certain continuous components. As a general guideline for which supporting theory remains to be developed, we expect that no components require smoothing if the parameter is sufficiently smooth at $P \in \mathcal{M}$ in the sense that $\Psi(P_m)$ tends to $\Psi(P)$ for any sequence $\{P_m \in \mathcal{M} : m = 1, 2, \dots\}$ for which the cumulative distribution of P_m tends to that of P uniformly as m tends to infinity. Alternatively, if the MLE P_n^* of P based on observations X_1, X_2, \dots, X_n from P is such that $\Psi(P_n^*)$ is a consistent estimator of $\Psi(P)$, no smoothing will generally be required. If, however, some regularization of the MLE is needed to ensure consistency (see, e.g., van der Laan, 1996), smoothing will usually be critical.

4.2 Selection of ϵ and λ values

When the pathwise derivative in (5) can be calculated analytically, the approximation method proposed only involves the smoothing parameter λ . The supporting theory clearly suggests choosing λ to be as small as possible. As we will illustrate in Section 5, in some cases there is little sensitivity to the choice of λ when (5) is used, and even a relatively large value of $\lambda > 0$ will yield stringent control of the approximation error.

Whenever the involved projection is not available in closed form or differentiation with respect to ϵ is too cumbersome to perform analytically, the secant line slope may be used to numerically approximate this analytic derivative. In such case, ϵ and λ must both be chosen, and more care is needed to ensure the reliability of the proposed procedure. The order of the limits in (5) and (6) suggests that we must select a small value of λ and even smaller value of ϵ . This was made more precise in Section 3, where it is prescribed to choose ϵ to be much smaller than λ^{2d_1} , where d_1 is the number of components of P over which smoothing is required. While this theoretical requirement may serve as a rough guide in practice, it does not provide a concrete means of selecting values for ϵ and λ . For this purpose, it may be useful to produce a matrix representing the value of

$$\frac{\Psi(P_{\epsilon, \lambda}^*) - \Psi(P)}{\epsilon}$$

as a function of ϵ and λ , both ranging over an exponential scale – for example, we could consider both ϵ and λ in the set $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, \dots\}$. We refer to the resulting display as an epsilon-lambda plot. As a convention, the y-axis is used to represent ϵ values while λ values are represented

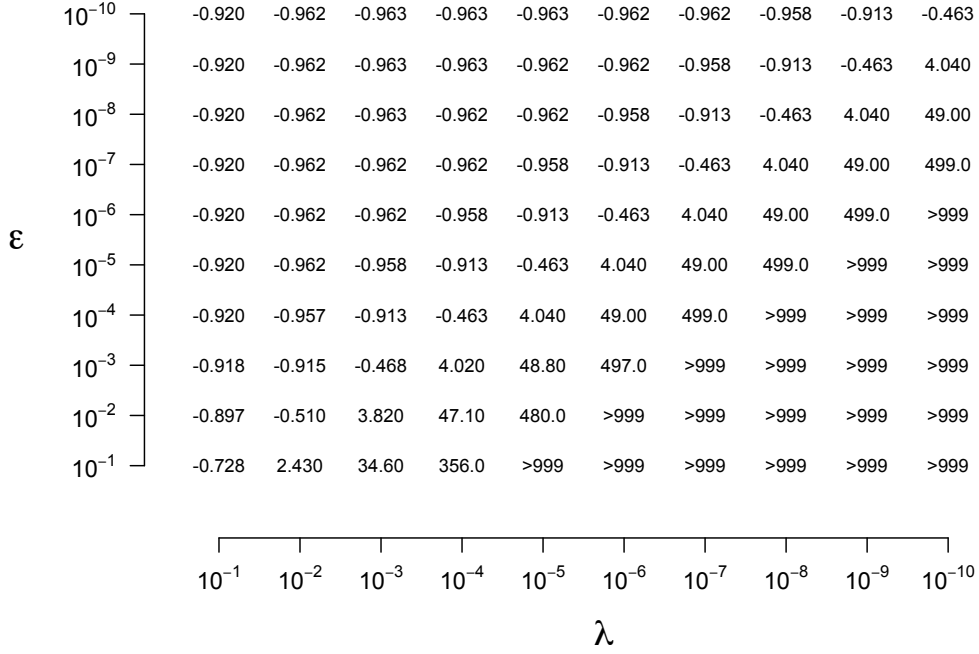


Figure 1: Epsilon-lambda plot of approximated values of the EIF using a secant line slope as a function of ϵ and λ in Example 1

on the x-axis. Our theoretical findings suggest that the right balance between ϵ and λ will be achieved in a possibly curvilinear triangular region nested in the upper left portion of the epsilon-lambda plot. In this triangular region, the secant line slope should be essentially constant. One practical means of selecting ϵ and λ would then consist of identifying this region visually by determining the quasi-triangular region in the upper left portion of the matrix over which the approximated EIF value is fixed up to a certain level of precision. As an illustration, without yet providing details regarding the specific parameter and model under consideration, we may scrutinize the epsilon-lambda plot arising in Example 1 from Section 5. This plot is provided as Figure 1 and clearly suggests that, up to three decimal points, the EIF value of interest is -0.963. This is indeed verified using theoretical calculations, as discussed in more detail in Section 5. The epsilon-lambda plot therefore may be a particularly useful tool for implementing the proposed approach for numerically approximating the EIF in practice.

4.3 Numeric computation of the model space projection

In implementing our proposal, the main challenge consists of operationalizing the optimization problem that characterizes the projection of the linear perturbation path $\{P_{\epsilon,\lambda} : 0 \leq \epsilon \leq 1\}$ onto the model space \mathcal{M} . An analytic – or nearly analytic – form can be found for the projection in many problems, including the illustrations provided in Section 5. In other problems, the optimization problem is less analytically tractable and a numeric approach may be needed.

A general strategy for numerically approximating the required projection is to instead consider the corresponding optimization problem over \mathcal{M}_m , where $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots \subseteq \mathcal{M}$ is a sequence of finite-dimensional submodels of \mathcal{M} such that $\cup_{m=1}^{\infty} \mathcal{M}_m = \mathcal{M}$. We illustrate this in the context of families of tilted densities, though many other parametrizations are possible. We note that any distribution Q dominated by P can be described as a tilted form $dQ(u) = \exp\{h(u)\}dP(u) / \int \exp\{h(w)\}dP(w)$ of P for some function $h \in \mathcal{H} := \mathcal{H}(\mathcal{M})$ in a function class determined by the model \mathcal{M} . Here, h characterizes the deviation of Q from P . It is often easier to determine suitable approximating finite-dimensional subspaces for \mathcal{H} than for \mathcal{M} . Suppose that $\{h_1, h_2, \dots\} \subseteq \mathcal{H}$ forms a basis for \mathcal{H} , and let \mathcal{H}_m denote the linear span of $\{h_1, h_2, \dots, h_m\}$. If P has density p relative to ν , the submodel \mathcal{M}_m implied by \mathcal{H}_m then consists of all distributions Q with density given by

$$\frac{dQ}{d\nu}(u) = \frac{\exp\left\{\sum_{j=1}^m \beta_j h_j(u)\right\}p(u)}{\int \exp\left\{\sum_{j=1}^m \beta_j h_j(w)\right\}p(w)\nu(dw)}$$

for some $\underline{\beta}_m := (\beta_1, \beta_2, \dots, \beta_m) \in \mathbb{R}^m$. The choice $\underline{\beta}_m = (0, 0, \dots, 0)$ leads to $Q = P$. Denoting by $P_{\epsilon,\lambda,m}^*$ the projection of $P_{\epsilon,\lambda}$ onto \mathcal{M}_m , this suggests that the corresponding optimizer $\underline{\beta}_m^*(\epsilon, \lambda)$ should be near zero for small ϵ since then $P_{\epsilon,\lambda}^* \approx P$. Thus, the search for the optimizer can be focused in a neighborhood surrounding the origin in \mathbb{R}^m . This simple observation can sometimes greatly accelerate the numerical optimization routine used. In practice, a sufficiently large m must be selected to ensure that the resulting approximation of the projection is accurate enough to ensure the validity of the numerical evaluation of the EIF based on (6). Up to an additive constant, the resulting objective function to maximize is

$$\mathcal{L}(\underline{\beta}_m) := \sum_{j=1}^m \beta_j \int h_j(u) dP_{\epsilon,\lambda}(u) - \log \int \exp\left\{\sum_{j=1}^m \beta_j h_j(w)\right\} p(w) \nu(dw) .$$

Since derivatives of $\mathcal{L}(\underline{\beta}_m)$ are easy to write down explicitly, many algorithms are available to solve

this optimization problem efficiently, including Newton’s method.

It may sometimes be useful to consider a stochastic version of this deterministic optimization problem. Specifically, we may generate a very large number of observations from $P_{\epsilon,\lambda}$ – this is often easy because $P_{\epsilon,\lambda}$ is no more than a mixture between P and $H_{x,\lambda}$ – and write the likelihood of the approximating finite-dimensional submodel based on these data. We are then faced with a standard parametric estimation problem, albeit one that may be high-dimensional. When a clever parametrization of the approximating submodel is used, it is often possible to employ standard statistical learning techniques, including regularization methods from the machine learning literature, using computationally efficient and stable off-the-shelf implementations. When adopting this approach, it appears critical to ensure that the size of the dataset generated is very large compared to the richness of the approximating submodel, since otherwise the variability resulting from this parametric estimation problem could limit our ability to achieve the required level of accuracy.

4.4 Construction of an efficient estimator

As emphasized earlier, knowledge of the EIF facilitates the construction of efficient estimators in infinite-dimensional models. For example, if \hat{P}_n is a consistent estimator of $P_0 \in \mathcal{M}$ based on independent draws X_1, X_2, \dots, X_n from P_0 , the corresponding one-step Newton-Raphson estimator, defined as

$$\psi_n^+ := \Psi(\hat{P}_n) + \frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(X_i) ,$$

is an efficient estimator of ψ_0 under certain regularity conditions. The one-step approach appears to be the constructive method most amenable to an implementation based on numerical approximations of the EIF. Indeed, if the analytic form of the EIF is not known, it suffices to numerically approximate the value of $\phi_{\hat{P}_n}(X_i)$ for each $i = 1, 2, \dots, n$, rather than the entire function $u \mapsto \phi_{\hat{P}_n}(u)$, in order to calculate ψ_n^+ . Thus, the procedure described in this paper can be used to approximate each of these n values. Nevertheless, when the projection step required to utilize the proposed representations of the EIF is computationally burdensome and the sample size n is large, computing each of these values may be challenging. One need not obtain an approximation of each $\phi_{\hat{P}_n}(X_i)$ if our objective is only to compute the one-step estimator ψ_n^+ – in this case it suffices to obtain an approximation of the empirical average $\frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(X_i)$. This simple observation is useful because a slight modification to

the representations of the EIF introduced in this paper yields a numerical procedure for approximating the required empirical average. Specifically, it is straightforward to adapt the proof of Theorem 1 to show that, under similar regularity conditions, if we define the linear perturbation $\hat{P}_{n,\epsilon,\lambda} := (1 - \epsilon)\hat{P}_n + \epsilon \frac{1}{n} \sum_{i=1}^n H_{X_i,\lambda}$ between \hat{P}_n and a uniform mixture of nearly degenerate distributions on $\{X_1\}, \{X_2\}, \dots, \{X_n\}$, it follows that

$$\frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(X_i) = \lim_{\lambda \rightarrow 0} \left. \frac{d}{d\epsilon} \Psi(\hat{P}_{n,\epsilon,\lambda}^*) \right|_{\epsilon=0} = \lim_{\lambda \rightarrow 0} \lim_{\epsilon \rightarrow 0} \frac{\Psi(\hat{P}_{n,\epsilon,\lambda}^*) - \Psi(\hat{P}_n)}{\epsilon}$$

with $\hat{P}_{n,\epsilon,\lambda}^* := \operatorname{argmax}_{P_1 \in \mathcal{M}(P)} \int \log \left\{ \frac{dP_1}{d\nu}(u) \right\} d\hat{P}_{n,\epsilon,\lambda}(u)$. As such, a numerical approximation of the one-step estimator can be computed in a single numerical step as

$$\Psi(\hat{P}_n) + \left. \frac{d}{d\epsilon} \Psi(\hat{P}_{n,\epsilon,\lambda}^*) \right|_{\epsilon=0} \approx \Psi(\hat{P}_n) + \frac{\Psi(\hat{P}_{n,\epsilon,\lambda}^*) - \Psi(\hat{P}_n)}{\epsilon}$$

for appropriately selected ϵ and λ values.

5 Illustration and numerical studies

To illustrate use of the representations presented above, we consider two particular examples in which the calculation of the EIF can be difficult for non-experts, whereas the approach proposed in this paper renders the problem straightforward. The technical conditions required for representations (5) and (6) to hold are satisfied in these examples with the distributions selected, although we do not include details of these verifications here.

5.1 Example 1: Average density value under known population mean

5.1.1 Background

Given a distribution P with Lebesgue density p , the average density value parameter is given by

$$\Psi(P) := E_P \{p(X)\} = \int p(u)^2 du .$$

Estimation and inference for the average density value has been extensively studied in the semiparametric efficiency literature (see, e.g., Bickel and Ritov, 1988). We use this parameter as our first illustration because it is simple to describe yet requires specialized knowledge to study using conven-

tional techniques. Suppose that \mathcal{M}_{NP} denotes the nonparametric model consisting of all univariate absolutely continuous distributions with finite-valued density. Suppose that $\mu \in \mathbb{R}$ is fixed and known, and denote by $\mathcal{M} \subset \mathcal{M}_{\text{NP}}$ the semiparametric model consisting of all distributions in \mathcal{M}_{NP} with mean μ . We wish to compute the EIF of Ψ relative to \mathcal{M} at a distribution $P \in \mathcal{M}$ evaluated at an observation value x .

The EIF $\phi_{\text{NP},P}$ of Ψ relative to the nonparametric model \mathcal{M}_{NP} evaluated at $P \in \mathcal{M}$ is given by $u \mapsto \phi_{\text{NP},P}(u) := 2 \{p(u) - \Psi(P)\}$ – it is rather straightforward to derive this analytic form from first principles. Observing that $\mathcal{M} = \{P \in \mathcal{M}_{\text{NP}} : \Theta(P) = 0\}$, where $\Theta(P) := \int u dP(u) - \mu$ is a pathwise differentiable parameter with EIF relative to \mathcal{M}_{NP} at $P \in \mathcal{M}$ given by $u \mapsto \varphi_P(u) := u - \mu$, Example 1 of Section 6.2 of Bickel et al. (1997) suggests that the EIF of Ψ relative to \mathcal{M} can be obtained as

$$\begin{aligned} u \mapsto \phi_P(u) &:= \phi_{\text{NP},P}(u) - \frac{\int \phi_{\text{NP},P}(w) \varphi_P(w) dP(w)}{\int \varphi_P(w)^2 dP(w)} \varphi_P(u) \\ &= 2 \left\{ p(u) - \Psi(P) - \frac{\int (w - \mu) p(w) dP(w)}{\int (w - \mu)^2 dP(w)} (u - \mu) \right\}. \end{aligned}$$

While the resulting analytic form of this EIF is relatively simple, its derivation hinges on specialized knowledge unlikely to be available to most practitioners. Use of our novel representation of the EIF provides an alternative approach that avoids the need for such knowledge, as highlighted below.

5.1.2 Implementation and results

To utilize our representation, we must understand how to project a given distribution $Q \in \mathcal{M}$, say with Lebesgue density q , into \mathcal{M} relative to the Kullback-Leibler divergence. Suppose that the support of Q has finite lower and upper limits a and b , respectively, satisfying that $a < \mu < b$. An application of the method of Lagrange multipliers yields that the maximizer in p of $\int \log p(u) dQ(u)$ over the class of all Lebesgue densities with mean μ is given by $q^*(u) := \{1 - \xi_0(u - \mu)\}^{-1} q(u)$, where $\xi_0 \in \mathbb{R}$ solves the equation

$$\int \left\{ \frac{u}{1 - (u - \mu)\xi} - \mu \right\} dQ(u) = 0 \quad (1)$$

in ξ and lies strictly between $(a - \mu)^{-1}$ and $(b - \mu)^{-1}$.

To compute $\phi_P(x)$ using the approach proposed in this paper, we must first construct the linear perturbation $P_{\epsilon,\lambda} := (1 - \epsilon)P + \epsilon H_{x,\lambda}$, where $H_{x,\lambda}$ is an absolutely continuous distribution that concentrates its mass on shrinking neighborhoods of the set $\{x\}$ as λ tends to zero. For example, we

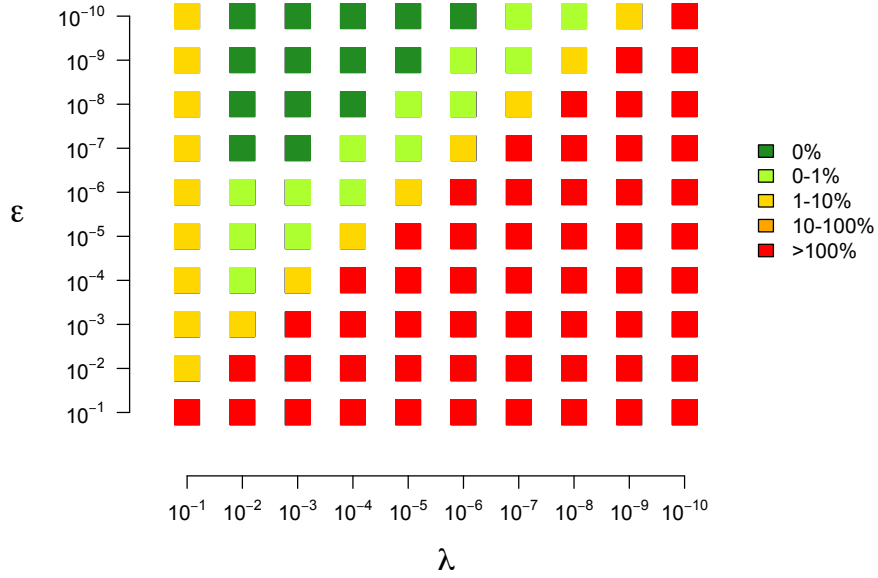


Figure 2: Absolute % error in the approximation of the EIF value using a secant line slope as a function of ϵ and λ in Example 1

may take $H_{x,\lambda}$ to be the uniform distribution on the interval $(x - \lambda, x + \lambda)$. The projection of $P_{\epsilon,\lambda}$ onto \mathcal{M} is then obtained as described in the preceding paragraph with $Q = P_{\epsilon,\lambda}$ – as such, it has a closed-form analytic expression up to the constant $\xi_0 = \xi_0(\epsilon, \lambda)$ that can be numerically solved. In the Supplementary Material, we study some properties of ξ_0 . We may then approximate $\phi_P(x)$ by the secant line slope

$$\phi_P(x) \approx \frac{\Psi(P_{\epsilon,\lambda}^*) - \Psi(P)}{\epsilon}.$$

We evaluated this procedure numerically for a particular distribution P and observation value x . Specifically, we took P to be the Beta distribution with parameters $\alpha = 3$ and $\beta = 5$, and evaluated our numerical procedure for approximating the true value of $\phi_P(0.6) \approx -0.963$. Figure 2 provides the percent error of our numerical approximation for various combination of values for ϵ and λ . This approximation is inaccurate if either ϵ is not small enough or if λ is too small relative to ϵ . For small λ and much smaller ϵ , the secant line slope approximates the true value of $\phi_P(x)$ with a relative error below 0.1%. This plot confirms what theory suggests regarding the choice of ϵ and λ . It also reaffirms the usefulness of the epsilon-lambda plot for selecting appropriate values of ϵ and λ .

5.2 Example 2: G-computation parameter under Markov structure

5.2.1 Background

We now consider a more complex parameter arising in the causal inference literature. Suppose that the data unit consists of the longitudinal observation $X := (L_0, A_0, \dots, L_K, A_K, L_{K+1}) \sim P_0$, where L_0, L_1, \dots, L_K is a sequence of measurements collected at $K+1$ distinct instances through time, L_{K+1} is the outcome of interest, and A_0, A_1, \dots, A_K are intervention indicators corresponding to each pre-outcome timepoint. For simplicity, we consider all treatment indicators to be binary. Let \mathcal{M}_{NP} be a nonparametric model. In practice, we may be interested in the covariate-adjusted, treatment-specific mean $\psi_0 := \Psi(P_0)$ corresponding to the intervention $(A_0, A_1, \dots, A_K) = (1, 1, \dots, 1)$. Here, for any given $P \in \mathcal{M}_{\text{NP}}$, the parameter value $\Psi(P)$ is defined explicitly as $E_P[m_{0,P}(L_0)]$ via the G-computation recursion

$$m_{j,P}(\bar{\ell}_j) := E_P [m_{j+1,P}(\bar{L}_{j+1}) \mid \bar{L}_j = \bar{\ell}_j, A_j = A_{j-1} = \dots = A_0 = 1]$$

for $j = K, K-1, \dots, 0$, where we have set $m_{K+1,P}(\bar{L}_{K+1}) := L_{K+1}$ (Robins, 1986). Here, for any vector $u := (u_0, u_1, \dots)$ we write $\bar{u}_k := (u_0, u_1, \dots, u_k)$. This parameter only depends on P through the conditional distribution $P_{j,1}$ of \bar{L}_{j+1} given \bar{L}_j and $A_0 = A_1 = \dots = A_j = 1$ for $j = 0, 1, \dots, K$, and the marginal distribution $P_{0,1}$ of L_0 . Under certain untestable causal assumptions, ψ_0 corresponds to the mean of the counterfactual outcome Y defined by an intervention setting all treatment nodes to one. With respect to \mathcal{M}_{NP} , or any model with restrictions only on the conditional distribution of A_j given \bar{A}_{j-1} and \bar{L}_j possibly for any $j \in \{0, 1, \dots, K\}$, the EIF of Ψ at P is known to be given by $\phi_{\text{NP},P} := \sum_{j=0}^{K+1} \phi_{j,\text{NP},P}$, where $\phi_{0,\text{NP},P}(x) := m_{0,P}(\ell_0) - \Psi(P)$ and

$$\phi_{j,\text{NP},P}(x) := \frac{a_0 a_1 \cdots a_{j-1}}{\prod_{r=0}^{j-1} P(A_r = 1 \mid \bar{L}_r = \bar{\ell}_r, A_0 = A_1 = \dots = A_{r-1} = 1)} \{m_{j,P}(\bar{\ell}_j) - m_{j-1,P}(\bar{\ell}_{j-1})\}$$

for $j = 1, 2, \dots, K+1$.

Let the model \mathcal{M} consist of the subset of distributions P in \mathcal{M}_{NP} such that, for each $j = 2, 3, \dots, K+1$, L_j and \bar{L}_{j-2} are independent given L_{j-1} and $A_{j-1} = A_{j-2} = \dots = A_0 = 1$ under P . For each $P \in \mathcal{M}$, we note that $m_{j,P}(\bar{\ell}_j) = m_{j,P}(\ell_j)$ for each j . The EIF of Ψ relative to \mathcal{M} at P is given by

$\phi_P := \sum_{j=0}^{K+1} \phi_{j,P}$, where $\phi_{0,P} = \phi_{0,NP,P}$ and $\phi_{j,P}$ is defined pointwise as

$$\begin{aligned} x \mapsto \phi_{j,P}(x) &:= E_P [\phi_{j,NP,P}(X) \mid L_j = \ell_j, L_{j-1} = \ell_{j-1}, \bar{A}_{j-1} = \bar{a}_{j-1}] \\ &\quad - E_P [\phi_{j,NP,P}(X) \mid L_{j-1} = \ell_{j-1}, \bar{A}_{j-1} = \bar{a}_{j-1}] \\ &= a_0 a_1 \cdots a_{j-1} \cdot T_j(P)(x) \cdot \{m_{j,P}(\ell_j) - m_{j-1,P}(\ell_{j-1})\} \end{aligned}$$

for $j = 1, 2, \dots, K+1$, and we use $T_j(P)(x)$ to denote

$$E_P \left[\frac{1}{\prod_{r=0}^{j-1} P(A_r = 1 \mid \bar{L}_r, A_0 = A_1 = \dots = A_{r-1} = 1)} \mid L_j = \ell_j, L_{j-1} = \ell_{j-1}, \bar{A}_{j-1} = \bar{a}_{j-1} \right].$$

Deriving this expression requires specialized knowledge and familiarity with efficiency theory for longitudinal structures. Furthermore, even given this analytic expression, the EIF may often be difficult to compute since it involves rather elaborate conditional expectations.

5.2.2 Implementation and results

As in the previous example, the main challenge is to understand how to project a given distribution Q into \mathcal{M} . Given a dominating measure ν , we denote the density function of Q with respect to ν as q . Furthermore, we denote by q_{L_j} the density of the conditional distribution of L_j given \bar{L}_{j-1} and \bar{A}_{j-1} , and by q_{A_j} the density of the conditional distribution of A_j given \bar{L}_j and \bar{A}_{j-1} . We also denote by $q_{L_{j,1}}$ the density q_{L_j} with $\bar{a}_{j-1} = (1, 1, \dots, 1)$. We use the same notational convention for any other candidate density p . Because for any candidate p we can write

$$\int \log p(u) dQ(u) = \sum_{j=0}^{K+1} \int \log p_{L_j}(\ell_j \mid \bar{\ell}_{j-1}, \bar{a}_{j-1}) dQ(u) + \sum_{j=0}^K \int \log p_{A_j}(a_j \mid \bar{\ell}_j, \bar{a}_{j-1}) dQ(u)$$

and \mathcal{M} can be written as a product model for the set of conditional distributions implied by the joint distribution, the required optimization problem can be performed separately for each conditional density. Because computing $\Psi(Q^*)$ does not require any component of Q^* beyond $q_{L_{j,1}}^*$ for $j = 0, 1, \dots, K+1$, we focus our attention on the corresponding optimization problems alone. Below, we denote by $\bar{q}(\ell_j, \ell_{j-1})$ the marginalized density $\iint \cdots \int q(\ell_0, 1, \ell_1, 1, \dots, \ell_{j-1}, 1, \ell_j) \nu(d\ell_0, d\ell_1, \dots, d\ell_{j-2})$. To find $q_{L_{j,1}}^*$ for $j = 2, 3, \dots, K+1$, we must maximize the criterion

$$L(p_{L_{j,1}}) := \iint \cdots \int \log p_{L_{j,1}}(\ell_j \mid \ell_{j-1}) q(\ell_0, 1, \ell_1, 1, \dots, \ell_{j-1}, 1, \ell_j) \nu(d\ell_0, d\ell_1, \dots, d\ell_j)$$

$$\begin{aligned}
&= \iint \log p_{L_j,1}(\ell_j \mid \ell_{j-1}) \bar{q}(\ell_{j-1}, \ell_j) \nu(d\ell_{j-1}, d\ell_j) \\
&= \iint \log p_{L_j,1}(\ell_j \mid \ell_{j-1}) \frac{\bar{q}(\ell_{j-1}, \ell_j)}{\int \bar{q}(\ell_{j-1}, \ell'_j) \nu(d\ell'_j)} \nu(d\ell_j) \int \bar{q}(\ell_{j-1}, \ell_j) \nu(d\ell_j) \nu(d\ell_{j-1})
\end{aligned}$$

over the class of candidate conditional densities that do not depend on $\bar{\ell}_{j-2}$, here represented by $p_{L_j,1}$. Since for each fixed ℓ_{j-1} the mapping $\ell_j \mapsto \bar{q}(\ell_{j-1}, \ell_j) / \int \bar{q}(\ell_{j-1}, \ell'_j) \nu(d\ell'_j)$ defines a proper conditional density, by Jensen's inequality, $L(p_{L_j,1})$ is maximized by

$$q_{L_j,1}^*(\ell_j \mid \ell_{j-1}) = \frac{\bar{q}(\ell_{j-1}, \ell_j)}{\int \bar{q}(\ell_{j-1}, \ell'_j) \nu(d\ell'_j)}.$$

It is easy to see that \mathcal{M} constrains neither $p_{L_1,1}$ nor p_{L_0} and therefore $q_{L_1,1}^* = p_{L_1,1}$ and $q_{L_0}^* = p_{L_0}$. Thus, in the context of a longitudinal data structure, the projection of any given distribution Q into a model only constrained by a Markov structure has an analytic closed-form.

As before, to compute $\phi_P(x)$ using the proposed representations of the EIF, we first construct the linear perturbation $P_{\epsilon,\lambda} := (1 - \epsilon)P + \epsilon H_{x,\lambda}$, where $H_{x,\lambda}$ is a distribution dominated by P and concentrating its mass in shrinking neighborhoods of the set $\{x\}$ as λ tends to zero. The projection $P_{\epsilon,\lambda}^*$ of $P_{\epsilon,\lambda}$ onto \mathcal{M} has an explicit form given in the preceding paragraph with $Q = P_{\epsilon,\lambda}$. As in Example 1, we may approximate $\phi_P(x)$ by the secant line slope $\{\Psi(P_{\epsilon,\lambda}^*) - \Psi(P)\}/\epsilon$ for small λ and even smaller ϵ . Because in this example $P_{\epsilon,\lambda}^*$ is available in closed form, $\phi_P(x)$ can alternatively be approximated by $\left. \frac{d}{d\epsilon} \Psi(P_{\epsilon,\lambda}^*) \right|_{\epsilon=0}$ for small λ .

For convenience, in our numerical evaluation of the EIF, we restricted our attention to a setting with $K = 2$ post-baseline time-points. We considered the joint distribution P of X defined in terms of the following conditional distributions. The baseline covariate L_0 has a discrete uniform distribution on the set $\{0, 1, 2, 3, 4\}$. Given $L_0 = \ell_0$, A_0 has a Bernoulli distribution with success probability $\text{expit}(-1 + 0.5\ell_0)$. Given $A_0 = a_0$ and $L_0 = \ell_0$, L_1 has a normal distribution with mean $3\ell_0 - 3a_0$ and variance 4. Given $L_1 = \ell_1$, $A_0 = a_0$ and $L_0 = \ell_0$, A_1 has a Bernoulli distribution with success probability $\text{expit}\{-5 + c_{10}(\ell_1) + a_0 + 0.5\ell_0\}$, where we define c_{10} to be the trimming function $u \mapsto -10 \cdot I_{(-\infty, -10)}(u) + u \cdot I_{[-10, +10]}(u) + 10 \cdot I_{(+10, +\infty)}(u)$. Given $A_1 = a_1$, $L_1 = \ell_1$, $A_0 = a_0$ and $L_0 = \ell_0$, Y has a Bernoulli distribution with success probability $\text{expit}\{-1 + 0.5c_{10}(\ell_1) - 0.5a_1 - a_0\}$. We evaluated the approximations of $\phi_P(x)$ based on either the secant line slope or the analytic pathwise derivative at various possible values of the realized data unit x . We report the absolute percent error for observation value $x := (0, 1, 2, 1, 1)$ using the secant line slope approach in Figure 3 and using the

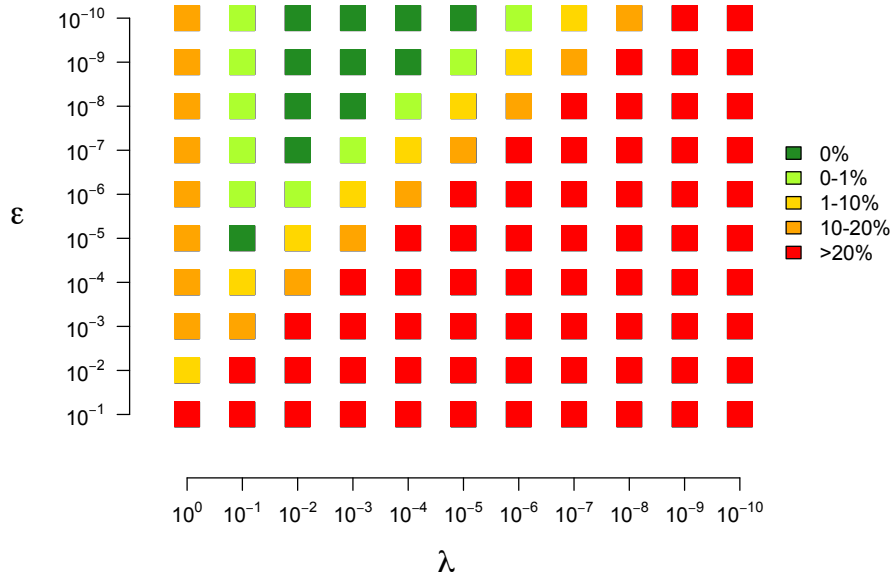


Figure 3: Absolute % error in the approximation of the EIF value using a secant line slope as a function of ϵ and λ in Example 2

analytic derivative approach in Figure 4. The pattern observed in Figure 3 is similar to that seen in Figure 2. In a triangular region contained in the upper left portion of the epsilon-lambda plot, the approximation provided by the secant line slope is very accurate. Outside of this region, that is, for inappropriate choices of ϵ and λ , the approximation can be poor. Thankfully, the epsilon-lambda plot provides an easy way of identifying these appropriate values. From Figure 4, we note that a high level of accuracy is achieved with a relatively large $\lambda > 0$. Thus, use of the analytic derivative essentially eliminates the careful selection of approximation parameters otherwise needed. Results for other observation values examined yielded similar patterns and are therefore not reported here.

6 Concluding remarks

The representations of the EIF we have presented in this paper suggest a natural strategy for numerically approximating the EIF. These representations hold in arbitrary models under mild regularity conditions. Use of these representations requires the ability to project a given distribution into the statistical mode – this is essentially no more than a maximum likelihood step that can be tackled by

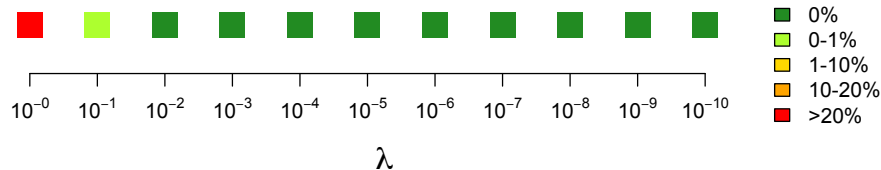


Figure 4: Absolute % error in the approximation of the EIF value using an analytic derivative as a function of ϵ and λ in Example 1

most practitioners. Most importantly, the involved work requires neither knowledge of efficiency theory nor familiarity with concepts from functional analysis or differential geometry. As such, these representations have the potential of democratizing the calculation of the EIF and thus the construction of efficient estimators in nonparametric and semiparametric models. Even for seasoned researchers in semiparametric and nonparametric theory, they provide an alternate means of tackling difficult problems, including those for which the EIF is either difficult or impossible to derive analytically.

In most problems, we anticipate the analytic work required to obtain the projection of the linear perturbation path onto the model space to be much simpler than that needed for the conventional tangent space approach. Nevertheless, this may still constitute a barrier for some practitioners. However, because the task of projecting onto the model space represents no more than an optimization problem, albeit an infinite-dimensional one, off-the-shelf computational tools may readily be used to circumvent most, if not all, analytic work otherwise required. This is particularly encouraging since strong computational skills are commonplace in statistics and data science. Furthermore, the numerical challenge will become increasingly surmountable as the capability of our computational devices continues to grow over time. It may therefore be particularly fruitful to invest additional energy into devising and studying broad numerical strategies for computerizing the calculation of the EIF based on the representations in this paper.

As with all methods that incorporate some level of automation and more readily lend themselves to use by non-specialists, there is a clear potential for misuse of the results we have presented. This appears to be an inevitable risk inherent to this type of proposal, and it equally applies to some of the most celebrated tools in current statistical practice, including the bootstrap. Deriving the EIF analytically undoubtedly remains the gold-standard approach and it should be preferred whenever possible since much information can be learned about the problem at hand from the analytic form of

the EIF. In particular, verification of the regularity conditions invoked in this paper can be difficult without prior analytic knowledge of the EIF. Nevertheless, the representations introduced in this paper have the potential of serving as an important new tool in the arsenal of statistical researchers and practitioners alike for performing semiparametric and nonparametric analyses. Devising algorithms for verifying the required regularity conditions in any given problem is an important avenue for future research.

We have noted that a distinct advantage of the representations we have provided is that once they have been used to compute the EIF of a certain parameter in a given statistical model, the EIF of any other parameter can be obtained without any additional work since the bulk of the work required is exclusively model-specific. Nevertheless, the involved computational work must be repeated for each observation value at which we wish to evaluate the EIF. In particular, this makes it difficult to approximate the entire EIF as a function, particularly in the case of continuous or longitudinal data units. While the one-step approach only requires the EIF at the observed data points, the implementation of other efficient estimators with potentially better properties, such as targeted minimum loss-based estimators (TMLE), generally requires the entire EIF. The representations presented in this paper are therefore not conducive to a computerized implementation of TMLE. There is promise that alternative representations may be better suited for this purpose – this is an area of active research.

Acknowledgments

MC gratefully acknowledges the support of NIAID grant 5UM1AI068635 and the Career Development Fund of the Department of Biostatistics at the University of Washington. MvdL gratefully acknowledges the support of NIAID grant 5R01AI074345.

References

- J.M. Begun, W.J. Hall, W.M. Huang, and J.A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, pages 432–452, 1983.
- P.J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.
- P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer, 1997.
- G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.
- C.E. Frangakis, T. Qian, Z. Wu, and I. Díaz. Deductive derivation and Turing-computerization of semiparametric efficient estimation (with discussion). *Biometrics*, 2015.
- J. Hájek. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(4):323–330, 1970.

- J. Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194, 1972.
- H. Ichimura and W.K. Newey. The influence function of semiparametric estimators. *arXiv preprint arXiv:1508.01378*, 2015.
- J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.
- Y.A. Koshevnik and B.Y. Levit. On a non-parametric analogue of the information matrix. *Theory of Probability & Its Applications*, 21(4):738–753, 1977.
- L. Le Cam. Limits of experiments. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 245–261, 1972.
- A.R. Luedtke, M. Carone, and M.J. van der Laan. A discussion of “Deductive derivation and Turing-computerization of semiparametric efficient estimation” by Frangakis et al. *Biometrics*, 2015.
- W.K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382, 1994.
- J. Pfanzagl. *Contributions to a general asymptotic statistical theory*. Springer, 1982.
- J M Robins. A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.
- C. Stein. Efficient nonparametric testing and estimation. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 187–195, 1956.
- M.J. van der Laan. Efficient estimation in the bivariate censoring model and repairing npml. *The Annals of Statistics*, 24(2):596–627, 1996.
- M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.
- M.J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer, 2011.
- M.J. van der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–23, 2004.
- A.W. van der Vaart. On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991.

Appendix

Proof of Theorem 2. If condition (a) holds, then the result is true because $\phi_{\epsilon,\lambda}^*$ is a score. We therefore consider the case where it does not hold. Since $\phi_{\epsilon,\lambda}^* \in T_{\mathcal{M}}(P_{\epsilon,\lambda}^*)$, there exists a sequence of one-dimensional regular parametric submodels $\mathcal{M}_{0,m} := \{P_{\gamma,m} : \gamma \in \Gamma_m\} \subset \mathcal{M}$ with $\Gamma_m \subset \mathbb{R}$ an interval containing zero and with score s_m for γ at $\gamma = 0$, $m = 1, 2, \dots$, such that

$$\|\phi_{\epsilon,\lambda}^* - s_m\|_{2,P_{\epsilon,\lambda}^*} \rightarrow 0$$

as m tends to infinity. For each $m = 1, 2, \dots$, we have that $\int s_m(u) dP_{\epsilon,\lambda}(u) = 0$. Because we can write

$$\begin{aligned} \left| \int \phi_{\epsilon,\lambda}^*(u) dP_{\epsilon,\lambda}(u) \right| &= \left| \int \{\phi_{\epsilon,\lambda}^*(u) - s_m(u)\} dP_{\epsilon,\lambda}(u) \right| = \left| \int \frac{dP_{\epsilon,\lambda}}{dP_{\epsilon,\lambda}^*}(u) \{\phi_{\epsilon,\lambda}^*(u) - s_m(u)\} dP_{\epsilon,\lambda}^*(u) \right| \\ &\leq \|\phi_{\epsilon,\lambda}^* - s_m\|_{2,P_{\epsilon,\lambda}^*} \left\| \frac{dP_{\epsilon,\lambda}}{dP_{\epsilon,\lambda}^*} \right\|_{2,P_{\epsilon,\lambda}^*} \end{aligned}$$

and under condition (b), there exists some $B \in (0, +\infty)$ such that $\|dP_{\epsilon,\lambda}/dP_{\epsilon,\lambda}^*\|_{2,P_{\epsilon,\lambda}^*} < B$ for sufficiently small ϵ and sufficiently smaller λ , it must be the case that $\int \phi_{\epsilon,\lambda}^*(u) dP_{\epsilon,\lambda}(u) = 0$. \square

Proof of Theorem 3. We first note that

$$\begin{aligned}
& \left| \int \phi_{\epsilon,\lambda}^*(u) d(H_{x,\lambda} - P)(u) - \phi_P(x) \right| \\
&= \left| \int \{\phi_{\epsilon,\lambda}^*(u) - \phi_P(u)\} dH_{x,\lambda}(u) + \int \phi_P(u) dH_{x,\lambda}(u) - \phi_P(x) - \int \phi_{\epsilon,\lambda}^*(u) dP(u) \right| \\
&\leq \left| \int \{\phi_{\epsilon,\lambda}^*(u) - \phi_P(u)\} dH_{x,\lambda}(u) \right| + \left| \int \phi_P(u) dH_{x,\lambda}(u) - \phi_P(x) \right| + \left| \int \phi_{\epsilon,\lambda}^*(u) dP(u) \right|
\end{aligned}$$

and because by assumption the second and third summands on the second line tend to zero as λ tends to zero, it suffices to study the first summand. We can bound this term by $\|\phi_{\epsilon,\lambda}^* - \phi_P\|_{\infty, \mathcal{S}_{x,\lambda}}$ and so, if condition (a) holds, the result follows immediately. Alternatively, we can write this term as

$$\begin{aligned}
\left| \int \{\phi_{\epsilon,\lambda}^*(u) - \phi_P(u)\} dH_{x,\lambda}(u) \right| &= \left| \int \frac{dH_{x,\lambda}}{dP}(u) \{\phi_{\epsilon,\lambda}^*(u) - \phi_P(u)\} dP(u) \right| \\
&\leq \|\phi_{\epsilon,\lambda}^* - \phi_P\|_{2,P} \left\| \frac{dH_{x,\lambda}}{dP} \right\|_{2,P}
\end{aligned}$$

and thus, if condition (b) holds, the result is also guaranteed to hold. \square

Proof of Theorem 4. Using that $P_{\epsilon,\lambda}^*$ is the maximizer of $Q \mapsto \int \log \left[\frac{dQ}{d\nu}(u) \right] dP_{\epsilon,\lambda}(u)$ over all $Q \in \mathcal{M}$, we note that

$$\begin{aligned}
0 &\geq \int \log \left[\frac{dP_{\epsilon,\lambda}^*}{dP}(u) \right] dP(u) = \int \log \left[\frac{dP_{\epsilon,\lambda}^*}{dP}(u) \right] d(P - P_{\epsilon,\lambda})(u) + \int \log \left[\frac{dP_{\epsilon,\lambda}^*}{dP}(u) \right] dP_{\epsilon,\lambda}(u) \\
&\geq \int \log \left[\frac{dP_{\epsilon,\lambda}^*}{dP}(u) \right] d(P - P_{\epsilon,\lambda})(u) \\
&= \epsilon \int \log \left[\frac{dP_{\epsilon,\lambda}^*}{dP}(u) \right] d(P - H_{x,\lambda})(u) \\
&= \epsilon \int \log \left[\frac{dP_{\epsilon,\lambda}^*}{dP}(u) \right] \left\{ 1 - \frac{dH_{x,\lambda}}{dP}(u) \right\} dP(u) .
\end{aligned}$$

Denoting for any pair $P_1 \ll P_2$ the function $u \mapsto \log \left[\frac{dP_1}{dP_2}(u) \right]$ by $L(P_1, P_2)$, this implies that

$$\begin{aligned}
\left| \int L(P_{\epsilon,\lambda}^*, P)(u) dP(u) \right| &\leq \epsilon \left| \int L(P_{\epsilon,\lambda}^*, P)(u) \left\{ 1 - \frac{dH_{x,\lambda}}{dP}(u) \right\} dP(u) \right| \\
&\leq \epsilon \|L(P_{\epsilon,\lambda}^*, P)\|_{2,P} \left\| 1 - \frac{dH_{x,\lambda}}{dP} \right\|_{2,P} \leq \epsilon \{1 + r(\lambda)\} \|L(P_{\epsilon,\lambda}^*, P)\|_{2,P} .
\end{aligned}$$

Provided $P_1 \ll P_2$, we have that $\int \{L(P_1, P_2)(u)\}^2 dP_2(u) \leq M \left| \int L(P_1, P_2)(u) dP_2(u) \right|$, where $M := M(P_1, P_2)$ depends on the supremum of the Radon-Nikodym of P_1 relative to P_2 (see, e.g., van der Laan et al., 2004). This allows us to write that

$$\left| \int L(P_{\epsilon,\lambda}^*, P)(u) dP(u) \right| \leq M \epsilon \{1 + r(\lambda)\} \left| \int L(P_{\epsilon,\lambda}^*, P)(u) dP(u) \right|^{\frac{1}{2}},$$

which directly implies that $\left| \int L(P_{\epsilon,\lambda}^*)(u) dP(u) \right| \leq M^2 \epsilon^2 \{1 + r(\lambda)\}^2$. If the Radon-Nikodym derivative

of P_1 relative to P_2 is bounded above by $0 < \zeta < +\infty$ over the support of P_2 , we can write that

$$\begin{aligned}
0 &\leq \left\| \frac{dP_1}{dP_2} - 1 \right\|_{2, P_2}^2 = \int \left\{ \frac{dP_1}{dP_2}(u) - 1 \right\}^2 dP_2(u) \\
&= \int \left\{ \sqrt{\frac{dP_1}{dP_2}(u)} - 1 \right\}^2 \left\{ \sqrt{\frac{dP_1}{dP_2}(u)} + 1 \right\}^2 dP_2(u) \\
&\leq K(\zeta) \int \left\{ \sqrt{\frac{dP_1}{dP_2}(u)} - 1 \right\}^2 dP_2(u) \leq K(\zeta) \left| \int L(P_1, P_2)(u) dP_2(u) \right|,
\end{aligned}$$

where $K(\zeta) := \zeta + 2\sqrt{\zeta} + 1$ and the last inequality is established using that $\log(u) \leq 2(\sqrt{u} - 1)$ for each $u > 0$. Thus, by assumption, we find that

$$R(P_{\epsilon, \lambda}^*, P) \leq C \left\| \frac{dP_{\epsilon, \lambda}^*}{dP} - 1 \right\|_{2, P}^2 \leq CK \left| \int L(P_{\epsilon, \lambda}^*, P)(u) dP(u) \right| \leq CKM^2\epsilon^2 \{1 + r(\lambda)\}^2$$

for small λ and sufficiently smaller ϵ , which directly establishes the theorem. \square